DOCUMENT RESUME

ED 270 464 TM 860 342

AUTHOR Doolittle, Allen E.
TITLE Gender-Based Differential Item Performance in Mathematics Achievement Items.
PUB DATE Apr 86
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Tests; Analysis of Variance; Estimation (Mathematics); Geometry; High Schools; High School Seniors; *Item Analysis; *Mathematics Achievement; *Mathematics Tests; Outcomes of Education; Research Design; *Sex Differences; *Test Items
IDENTIFIERS *ACT Assessment; *Differential Item Performance

ABSTRACT

A procedure for the detection of differential item performance (DIP) is used to investigate the relationships between characteristics of mathematics achievement items and gender differences in performance. Eight randomly equivalent samples of high school seniors were each given a unique form of the ACT Assessment Mathematics Usage Test (ACTM). Students without requisite math courses were deleted from the samples to control the possible confounding effect of differences in instruction at the high school level. Based on the remaining students, signed measures of DIP were obtained for each item in the eight ACTM forms. These DIP estimates were then investigated in a six by eight (item category by firm) experimental design. Using analysis of variance (ANOVA) procedures, a significant item category effect was found indicating a relationship between item characteristics and gender-based DIP. Follow-up analyses suggested that geometry and mathematics reasoning items had the largest negative impact on female examinees and more algorithmic, computation-oriented items were relatively easier for females. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270464

Gender-Based Differential Item Performance in Mathematics Achievement Items

Allen E. Doolittle

The American College Testing Program

P.O. Box 168

Iowa City, IA 52243

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A.E. Doolittle

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the American Educational Research Association.

San Francisco, April 1986

TM 860 342

Abstract

A procedure for the detection of differential item performance (DIP) is used to investigate the relationships between characteristics of mathematics achievement items and gender differences in performance. Eight randomly equivalent samples of high school seniors were each given a unique form of the ACT Assessment Mathematics Usage Test (ACTM). Students without requisite math courses were deleted from the samples to control the possible confounding effect of differences in instruction at the high school level. Based on the remaining students, signed measures of DIP were obtained for each item in the eight ACTM forms. These DIP estimates were then investigated in a 6 X 8 (item category by form) experimental design. Using ANOVA procedures, a significant item category effect was found indicating a relationship between item characteristics and gender-based DIP. Follow-up analyses suggested that Geometry and mathematics reasoning items had the largest negative impact on female examinees and more algorithmic, computation-oriented items were relatively easier for females.

Gender-Based Differential Item Performance in Mathematics Achievement Items

In recent years, test publishers and others in the measurement community have given increasing attention to procedures used to detect items that perform differentially for different groups. Differential item performance (DIP) is observed if, given examinees of equal abilities, the probability of answering the item correctly is related to group membership (Shepard, Camilli, and Averill, 1981; Petersen, 1980).

Procedures for detecting DIP may be applied in at least two ways. First, they may be employed as tools for screening items prior to test construction. Although there may be situations where this approach is helpful, it is not always practical. For some testing programs it is often very difficult to obtain adequate samples of examinees from relevant subgroups to systematically investigate for DIP in tryout items. A second use of DIP indices in test development is in investigating items for characteristics that might lead to a better understanding of differential performance. This second approach is followed by this study in an investigation of gender-based DIP in mathematics achievement items.

Research has shown that male high school students as a group

perform better than female high school students on mathematics achievement tests (Benbow and Stanley, 1982; Clark and Grandy, 1984; Fennema and Carpenter, 1981). A possible explanation is that male students typically receive more and/or a higher level of instruction in mathematics than do females (Fennema and Sherman, 1977). As a consequence, one might expect that instances of differential item performance in the form of an instructional effect against females might exist in mathematics achievement tests. DIP might be shown to exist for a higher level mathematics item if one group of students has been appropriately instructed in that concept and another group of students has not.

An earlier study (Doolittle, 1984), using data from one national administration of the ACT Assessment Mathematics Usage Test (ACTM), investigated the plausibility of a differential instruction interpretation of DIP in a situation where gender differences in mathematics background were known to exist. In the study, an index suggested by Linn and Harnisch (1981) was used to detect differentially performing items in six separate analyses. The analyses were based on comparisons of the different subgroups defined by various combinations of gender and academic background taken from the total sample.

The results provided support for the seemingly self-evident notion that differences in instructional background have a strong influence on mathematics achievement. However, the results did

not support the notion of gender-based DIP in mathematics achievement as a clear consequence of differences in instructional background. As predicted, more items were found with significant levels of DIP when the groups were defined by differences in instructional background than when they were defined by gender. But, contrary to the hypothesis that gender was simply a surrogate for level of mathematics instruction, for many items the direction of the DIP was often different for females than it was for the low instruction group. In other words, items that tended to work to the relative disadvantage of females were often found to disproportionately favor the low instruction groups, and vice versa.

The measure of instructional background used in the 1984 study was the number of semesters of mathematics instruction received in high school. Those in the sample who reported at least six semesters of mathematics (in an eight-semester high school career) were considered the high background group and those with less than six semesters were considered the low background group. A problem with this measure was that, although it was perhaps a reasonable measure of quantity of mathematics instruction, it said nothing about the type or quality of instruction. There could be substantial differences in the instructional backgrounds of students having the same number of mathematics courses to their credit.

This led to a second study involving a different form of the ACTM, a different sample of examinees, and what was perceived as an improved measure of instructional background (Doolittle, 1985). In addition to a simple count of the number of high school mathematics courses taken by the examinees, an indicator of the level of preparation was also taken into account by the instruction background index used in the study. Students reporting either eight semesters of high school math or participation in accelerated math, or both, were categorized as having a high level of mathematics background. Those who did not meet either of these criteria were considered the low background group. Despite the differences between the two pilot studies, the 1985 study generally confirmed the results of the 1984 study--there seemed to be a substantial gender effect that could not be explained by instructional differences at the secondary school level.

In addition, there seemed to be certain categories of ACTM items that differentially favored one group or another. Arithmetic and Algebraic Reasoning items (word problems) tended to favor "low background" students and Intermediate Algebra items tended to favor the "high background" group. When background level was controlled by the utilized measures of instruction, Geometry as well as Arithmetic and Algebraic Reasoning items were found to favor male examinees; other ACTM item categories tended

to favor females.

The results of the 1984 and 1985 studies suggested several conclusions.

1. Gender-based DIP that is not attributable to differences in instruction may exist in mathematics achievement items.
2. Differential item performance can be predicted based upon characteristics of the items and the examinees.
3. The Linn and Harnisch index has a reasonable degree of stability in situations such as these as demonstrated by the similarity of study results.

The present study was derived in part from these earlier observations. The primary objective was to build upon previous research to determine the existence of gender-based DIP in math achievement items when the possible confounding effect of differential instruction is minimized. Previous research efforts struggled with the problem of assessing level of instruction. To the extent that the approaches used for measuring background could be questioned, so too could the conclusions of the research. To help clarify the investigation, this study relied on more specific background data to select male and female students with essentially equivalent training in high school mathematics.

A second objective of the present research was to investigate specific item content as it relates to gender-based DIP. An experimental design approach was followed, somewhat similar to the

approach used by Schmeiser and Ferguson (1978) and Schmeiser (1983). Multiple forms of the ACTM were used to gather DIP information on a large group of items previously classified into six content categories. The results of the previous studies suggested that several of these content categories might be relevant to an understanding of gender-based DIP in mathematics. When mathematics background is controlled at a level where all students have had the requisite training, Geometry and Arithmetic and Algebraic Reasoning items were predicted to favor male examinees. On the other hand, algebra and calculation-oriented items (Intermediate Algebra, Number and Numeration Systems, and Arithmetic and Algebraic Operations items) were predicted to relatively favor females.

METHODOLOGY

Data Source

The data for this research was drawn from a sample of college-bound, high school seniors from the October 1985 administration of the ACT Assessment Mathematics Usage Test (ACTM). Eight forms of the ACTM were administered to the students in a spiraled fashion thus creating eight randomly equivalent samples of students. Only those students with mathematics background in certain mathematics courses were considered. The final data sets were eight randomly equivalent samples of 1,300-1,400 students apiece (see Table 1). Approximately 55% of the students were female.

Insert Table 1 about here

The Instrument

The ACT Assessment program contains educational achievement tests in four content areas, one of which is Mathematics Usage (ACTM). The ACTM is a 40-item, 50-minute measure of mathematical reasoning ability. It emphasizes the solution of practical, quantitative problems that are encountered in many postsecondary programs and includes a sampling of mathematical techniques covered in high school courses. The test emphasizes quantitative

reasoning rather than memorization of formulas, knowledge of techniques, or computational skill. In general, the mathematical skills required for the test involve proficiencies emphasized in high school plane geometry and first- and second-year algebra. Six types of items are included in the test and are described below.

1. Arithmetic and Algebraic Operations (AAO). The items in this category explicitly describe operations to be performed by the student. The operations include manipulating and simplifying expressions containing arithmetic or algebraic fractions, performing basic operations in polynomials, solving linear equations in one unknown, and performing operations on signed numbers. Four items of this type are included on each form of the ACTM.
2. Arithmetic and Algebraic Reasoning (AAR). These word problems present practical situations in which algebraic and/or arithmetic reasoning is required. The problems require the student to interpret the question and to either solve the problem or find an approach to its solution. Fourteen AAR items are included on each form of the ACTM.

3. Geometry (G). The items in this category cover such topics as measurement of lines and plane surfaces, properties of polygons, the Pythagorean theorem, and relationships involving circles. Both formal and applied problems are included. Each form of the ACTM includes eight G items.
4. Intermediate Algebra (IA). The items in this category cover such topics as dependence and variation of quantities related by specific formulas, arithmetic and geometric series, simultaneous equations, inequalities, exponents, radicals, graphs of equations, and quadratic equations. Eight IA items are included in the ACTM.
5. Number and Numeration Concepts (NNS). The items in this category cover such topics as rational and irrational numbers, set properties and operations, scientific notation, prime and composite numbers, numeration systems with bases other than 10, and absolute value. Four NNS items are included on each form of the ACTM.
6. Advanced Topics (AT). The items in this category cover such topics as trigonometric functions, permutations and combinations, probability, statistics, and logic. Only simple applications of the skills implied by these topics are tested. Each form of the ACTM includes two AT items.

Measures of Instructional Background

As part of the registration process for the ACT Assessment, examinees were asked to indicate whether or not they have taken specific mathematics courses. Examinees were included if they reported having completed a course in Geometry, Advanced Algebra or Algebra II, and either or both Trigonometry and Advanced Math (includes pre-Calculus). Approximately 40% of the college bound seniors met this requirement. Mean performance, by form, of the selected students in contrast to the total group of examinees is shown in Table 2.

Insert Table 2 about here

Index of Differential Item Performance

An index suggested by Linn and Harnish (1981) was used to measure DIP. Although this measure is based on the three-parameter logistic model (Birnbaum, 1968), it may be viewed as a "small sample" alternative to some of the more theoretically preferred IRT-based indices. Applicability to small samples is considered to be a major advantage of this measure, because it is not uncommon for a subgroup to be small, even when the overall size of the data set is reasonably large.

Like most other indices of DIP, the Linn and Harnisch index is a relative, not an absolute measure of DIP. That is, the index

assumes that the total test score is an unbiased measure of ability or achievement. With this assumption, DIP exists when the performance of an item for a particular group is not in line with the performance of the total group.

To calculate the Linn and Harnisch index, the item and ability parameters of the three-parameter logistic model are estimated for the total sample. A target group (females in this study) is then separated from the rest of the sample. The difference is taken between each target group examinee's probability of correct response to an item, obtained from the model, and the examinee's actual response to the item (1=correct; 0=incorrect). The index is this difference, standardized and averaged over all members of the target group. This index is considered a signed index. That is, the sign indicates the direction of the DIP. As calculated here, negative values represent DIP against the target group and positive values represent DIP favoring the target group. Previous research has shown the Linn and Harnisch measure to be a reliable index and to be substantially correlated with other, perhaps more common, measures of DIP (Doolittle, 1983).

Design and Analysis

A random replications design was used to investigate the effect of mathematics item category on gender-based DIP. Item

category was considered a fixed effect and test form was considered a random effect. All six ACTM item categories were crossed with the eight unique forms, used essentially as replications, creating 48 distinct cells. Individual items were nested within form and item category.

Separately for each form, the Linn and Harnisch (1981) procedure was used to estimate DIP indices for each of the 40 items. Negative values of the index represented DIP favoring males; positive values represented DIP favoring females. The analysis was unweighted with the observed score in each cell as the signed, mean DIP index for the items in the cell. Analysis of variance procedures were used to determine whether or not there was a significant item category effect on gender-based DIP.

Results

Table 3 shows the means of the DIP indices for each item category and each form. Mean index values for the six item categories, averaged across all forms, are also presented. Inspection of the means across forms suggests some stable patterns. Geometry (G) and Arithmetic and Algebraic Reasoning (AAR) items have negative means for each of the eight forms. On the other hand, Intermediate Algebra (IA) and Arithmetic and Algebraic Operations (AAO) items have predominantly positive means across the forms.

Insert Table 3 about here

To determine the significance of the observed mean differences in DIP indices for the item categories, an unweighted random replications analysis of variance was performed. Item category was considered a fixed effect and form was considered random. The results of the analysis are summarized in Table 4.

Insert Table 4 about here

Only the item category main effect was found to be significant. Since the test forms were constructed to be as equivalent as possible based upon detailed specifications, it was not surprising that the form main effect and the category by form interaction were not significant.

The Scheffe procedure was used to test for differences among the means associated with the item category main effect. The results of this analysis are represented in Table 5. These results suggest that the mean DIP index for Geometry items is significantly lower than the NNS, IA, and AAO item category means and that the AAR mean is significantly lower than the IA and AAO means. In other words, Geometry and AAR (word problem) items tend to be relatively more difficult for female examinees and,

conversely, IA and AAO items are relatively less difficult for females.

Insert Table 5 about here

Discussion

The results of this study suggest that some gender-based differential item performance exists in mathematics achievement items and that the DIP is not a simple consequence of differential instruction at the high school level. However, the cause or causes of gender differences in performance on certain items is not clear. Two possibilities come to mind. Perhaps the ACTM is "biased" in the sense of unfairly measuring performance on certain items, or it may be that group differences in instruction or background have been so well established prior to high school that balancing on the basis of the high school curriculum is not enough of a control.

To think in terms of the first possibility, that is to conceive of the ACTM as "biased" in the sense of unfairness, does not seem to be particularly useful. Each form of the ACTM is carefully assembled from specifications directly tied to high school mathematics curricula. Even with equivalent high school course backgrounds, it seems most likely that differences in the learning of mathematics concepts do exist among high school seniors and that it is these differences that are being assessed by items in the ACT Mathematics Usage Test.

If DIP is not due to unfairly biased measurement by test items, it may be that the differential background hypothesis is

still tenable. However, background may need to be considered as broader than just a particular set of high school mathematics courses. This line of thinking would support the Fennema and Sherman (1977) position that attitudes, extra-curricular activities and a wide array of sociocultural factors that accumulate throughout the lives of individuals must be considered for a reasonably complete description of background.

The primary feature of this study was the identification of certain item characteristics that are related to gender-based differential item performance. Consistent patterns of DIP by item classification were observed which suggest a number of hypotheses. The fact that Geometry items, many of which contained diagrams, relatively favored male examinees might support the position that male examinees, as a group, tend to have developed certain spatial skills to a greater degree than females (Benbow & Stanley, 1982; Maccoby & Jacklin, 1974). Other outcomes might suggest that males have developed relatively stronger mathematics reasoning skills (measured partially by AAR items) and that females have developed relatively stronger algorithmic or computational skills (AAO, IA, and NNS items). Perhaps the most direct support for this latter hypothesis is the relative performance of each group on the Arithmetic and Algebraic Operations (AAO) and Arithmetic and Algebraic Reasoning (AAR) items. Both classifications require essentially the same

knowledge of mathematics concepts. The primary difference between the item types is in the context. The AAOs involve the manipulation, simplification, or solution of explicitly described operations. AARs, on the other hand, are verbal presentations of practical situations that require the examinee to determine an appropriate solution strategy before carrying out the necessary operations.

The approach taken by this study was an application of DIP methodology, within the context of an experimental design, to investigate the relationship between item characteristics and group performance. This use of a DIP procedure has provided useful information about the relative performance of male and female examinees on specific categories of ACT Assessment mathematics items. Perhaps in conjunction with future research that might focus on different dimensions of math achievement items, the results of this study should be useful in understanding more about the nature of differential item performance and in providing direction for instructional programs.

References

Benbow, C.P., & Stanley, J.C. (1982). Consequences in high school and college of sex differences in mathematical reasoning ability: a longitudinal perspective. American Educational Research Journal, 19, 598-622.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical Theories of Mental Test Scores, Reading, MA: Addison-Wesley Publishing Co., Inc., 404-405.

Clark, M.J., & Grandy, J. (1984). Sex differences in the academic performance of Scholastic Aptitude Test takers (College Entrance Examination Board Report 84-8; ETS Research Bulletin 84-43). New York: College Entrance Examination Board.

Doolittle, A.E. (1983, April). The reliability of measuring differential item performance. Paper presented at the annual meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED 234 061.)

Doolittle, A.E. (1984), April). Interpretation of differential item performance accompanied by gender differences in academic background. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 247 237.)

Doolittle, A.E. (1985, April). Understanding differential item performance as a consequence of gender differences in academic background. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Fennema, E., & Carpenter, T.P. (1981). Sex-related differences in mathematics: Results from the National Assessment. Mathematics Teacher, 74, 554-559.

Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. American Educational Research Journal, 14, 51-71.

Linn, R.L., & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. J. of Educational Measurement, 18, 109-118.

Maccoby, E., & Jacklin, C. (1974). Psychology of Sex Differences. Palo Alto, CA: Stanford University Press.

Petersen, N.S. (1980). Bias in the selection rule--bias in the test. InL.J. Th. van der Kamp, W.F. Langerak, and D.N.M. de Gruijter (Ed.), Psychometrics for Educational Debates. John Wiley & Sons, Ltd.

Schmeiser, C. B. (1983). Differences between black and white examinee performance on the ACT Assessment examination as a function of the racial orientation of test content. Doctoral dissertation, The University of Iowa.

Schmeiser, C.B., & Ferguson, R.L. (1978). Performance of black and white students on test materials containing content based on black and white cultures. J. of Educational Measurement, 15(3), 193-200.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. J. of Educational Statistics, 6, 317-375.

Table 1

Sampled Students by Gender Classification

	Form							
	A	B	C	D	E	F	G	H
Male	616	688	632	610	650	656	649	619
Female	775	725	691	711	735	704	734	718
Total	1,391	1,413	1,323	1,321	1,385	1,360	1,383	1,337

Table 2

Mean Performance by Form of Sampled Students

	Form							
	A	B	C	D	E	F	G	H
Sampled students								
Males	25.7	26.4	27.6	27.5	26.5	27.9	27.4	26.3
Females	22.7	23.9	25.5	24.4	24.4	25.6	24.4	23.8
Total	24.0	25.1	26.5	25.9	25.4	26.7	25.8	25.0
All students	20.4	21.1	21.6	21.3	21.3	21.9	21.3	20.9

Table 3

Means of DIP Indices

Category (items per form)	Form A	B	C	D	E	F	G	H	Total (All Forms)
AAO(4)	.01	.00	.05	.05	-.02	.03	.02	.04	.0224
AAR(14)	-.02	-.03	-.04	-.02	-.01	-.02	-.02	-.01	-.0206
IA(8)	.01	.02	.04	.03	.03	.03	.00	.02	.0211
G(8)	-.03	-.03	-.04	-.05	-.04	-.05	-.04	-.01	-.0370
NNS(4)	-.01	.02	.02	.01	.00	-.03	.01	.00	.0038
AT(2)	-.02	-.02	.04	-.02	-.03	.00	-.02	-.06	-.0172

Table 4

ANOVA Summary Table: Random Replications Analysis (Unweighted)

Source	SS	df	MS	F	F prob.
Category	0.0231	5	0.0046	18.135	.001
Form (Replications)	0.0058	7	0.0008		
Category X Form	0.0089	35	0.0003		
Total	0.0379	47			

Table 5

Significant Differences in Mean DIP Indices for Item Categories

(Scheffe follow-up procedure: overall = 0.10)

		Category					
Category	Mean	G	AAR	AT	NNS	IA	AAO
G	-.0370						
AAR	-.0206						
AT	-.0272						
NNS	.0038	*					
IA	.0211	*	*				
AAO	.0224	*	*				

Note: * denotes a significant difference in item category means.